High density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis

Published in final edited form as:
Nat Genet. 2015 February ; 47(2): 172–179. doi:10.1038/ng.3176.

High density mapping of the MHC identifies a shared role for *HLA-DRB1*01:03* in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis

Philippe Goyette[1,*], **Gabrielle Boucher**[1,*], **Dermot Mallon**[2,3], **Eva Ellinghaus**[4], **Luke Jostins**[5,6], **Hailiang Huang**[7,8], **Stephan Ripke**[7,8], **Elena S Gusareva**[9,10], **Vito Annese**[11,12], **Stephen L Hauser**[13], **Jorge R Oksenberg**[13], **Ingo Thomsen**[4], **Stephen Leslie**[14,15], **International IBD Genetics Consortium**[16], **Mark J Daly**[7,8], **Kristel Van Steen**[9,10], **Richard H Duerr**[17,18], **Jeffrey C Barrett**[19], **Dermot PB McGovern**[20], **L Philip Schumm**[21], **James A Traherne**[22,23], **Mary N Carrington**[24,25], **Vasilis Kosmoliaptsis**[2,3], **Tom H Karlsen**[26,27,28,^], **Andre Franke**[4,^], and **John D Rioux**[1,29,^]

[1]Research Center, Montreal Heart Institute, Montréal, Québec, Canada

[2]Department of Surgery, University of Cambridge, Cambridge, UK

[3]National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre, Cambridge, UK

[4]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany

[5]Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK

[6]Christ Church, University of Oxford, St Aldates, UK

[7]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

[8]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[9]Systems and Modeling Unit, Montefiore Institute, University of Liege, Liege, Belgium

[10]Bioinformatics and Modeling, Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA-R) Research Center, University of Liege, Liege, Belgium

Correspondence and requests for materials should be addressed to J.D.R. (John.david.rioux@umontreal.ca).
*These authors contributed equally to this work.
^These authors jointly supervised this work

URLs
Additional data on heterogeneity of effects for associated alleles: http://www.medgeni.org/goyette_nature_gen_2015 EBI sequence database: ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/

Author Contributions
J.D.R., M.J.D., K.V.S., V.K., T.H.K., A.F. jointly supervised research. J.D.R., P.G., G.B., R.H.D., J.C.B., D.P.B.M., J.A.T., M.N.C., V.K., A.F. conceived and designed the experiments. P.G., G.B., D.M., L.J. performed statistical analysis. P.G., G.B., L.J., E.S.G. analyzed the data. V.A., S.L.H., J.R.O., I.T., S.L., L.P.S. contributed reagents/materials/analysis tools. P.G., G.B., E.E., H.H., S.R. performed data QC and imputation. J.D.R., P.G., G.B., D.M., D.P.B.M., V.K., T.H.K., A.F. wrote the paper. All authors read and approved the final manuscript before submission

Competing Financial Interests
The authors declare no competing financial interests.

[11]Unit of Gastroenterology, Istituto di Ricovero e Cura a Carattere Scientifico-Casa Sollievo della Sofferenza (IRCCS-CSS) Hospital, San Giovanni Rotondo, Italy

[12]Unit of Gastroenterology SOD2, Azienda Ospedaliero Universitaria (AOU) Careggi, Florence, Italy

[13]Department of Neurology, University of California San Francisco, San Francisco, California, USA

[14]Murdoch Childrens Research Institute, Parkville, Victoria, Australia

[15]Department of Mathematics and Statistics, University of Melbourne, Melbourne, Victoria, Australia

[16]A full list of members and affiliations appears at the end of the paper

[17]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA

[18]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA

[19]Wellcome Trust Sanger Institute, Hinxton (Cambridge), UK

[20]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA

[21]Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA

[22]Cambridge Institute for Medical Research, Cambridge, UK

[23]Department of Pathology, University of Cambridge, Cambridge, UK

[24]Cancer and Inflammation Program, Laboratory of Experimental Immunology, Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, Frederick, Maryland, USA

[25]Ragon Institute of MGH, MIT and Harvard, Cambridge, Massachusetts, USA

[26]Research Institute of Internal Medicine, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway

[27]Norwegian PSC Research Center, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway

[28]K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[29]Faculté de Médecine, Université de Montréal, Montréal, Québec, Canada

Abstract

Genome-wide association studies of the related chronic inflammatory bowel diseases (IBD) known as Crohn's disease and ulcerative colitis have shown strong evidence of association to the major histocompatibility complex (MHC). This region encodes a large number of immunological candidates, including the antigen-presenting classical HLA molecules[1]. Studies in IBD have indicated that multiple independent associations exist at HLA and non-HLA genes, but lacked the

statistical power to define the architecture of association and causal alleles[2,3]. To address this, we performed high-density SNP typing of the MHC in >32,000 patients with IBD, implicating multiple HLA alleles, with a primary role for HLA-DRB1*01:03 in both Crohn's disease and ulcerative colitis. Significant differences were observed between these diseases, including a predominant role of class II HLA variants and heterozygous advantage observed in ulcerative colitis, suggesting an important role of the adaptive immune response to the colonic environment in the pathogenesis of IBD.

Meta-analyses of genome-wide association studies (GWAS) have recently shown that Crohn's disease (CD) (MIM266600) and ulcerative colitis (UC) (MIM191390) share the majority of the 163 known genetic risk factors for IBD, with the MHC being one of the notable exceptions[4]. Data from these GWAS, however, have had insufficient variant density to define the association signals within the MHC. Targeted studies of IBD with higher variant density within the MHC region but with modest sample sizes have indicated that multiple independent associations are likely to exist at human leukocyte antigen (HLA) genes and non-HLA genes, with the most consistent associations being to HLA class II loci, mainly *HLA-DRB1* and *HLA-DQB1*, with some reports of association at the *HLA-C* class I locus and potentially also at non-HLA genes[2,3,5–8]. In the current study we generated high quality genotypes for 7,406 SNPs within the MHC region on a total of 18,405 patients with CD, 14,308 patients with UC and 34,241 controls subjects. Using this SNP data, we imputed and benchmarked the genetic variation within the class I (*HLA-B*, *HLA-C*, and *HLA-A*) and class II (*HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1* and *HLA-DPB1*) HLA genes at the level of classical HLA alleles and amino acid positions (please refer to the Online Methods).

As a first step to defining the nature of the association to CD and UC within the MHC, we performed univariate analyses of the SNPs, classical HLA alleles, and HLA amino acids. These analyses revealed a very large number of variants across the MHC region with significant association to these phenotypes (Fig. 1), with major peaks of association centered in and around the classical HLA genes, suggesting a role for classical HLA alleles in CD and UC risk. This observation is consistent with gene-based analyses, which show strong association at the HLA genes for both UC and CD (e.g. $P<1\times10^{-300}$ for *HLA-DRB1* in UC) (**Supplementary Table 1**). In particular, these analyses demonstrated a role of *HLA-DRB1* that cannot be attributed to other HLA genes, with evidence of residual association in class I and class II regions (**Supplementary Table 1**). In order to be more quantitative, we calculated the variance explained by the class I and class II alleles. Whereas the contribution of class I and class II alleles are relatively equivalent in CD, not only is the overall impact of HLA on disease risk greater in UC, but the alleles in the class II region have nearly three-fold greater impact than class I alleles (Fig. 2). Moreover, these analyses have revealed that classical HLA alleles explain three- to ten-fold more of the disease variance than that explained by the index SNPs that were previously identified (~3% vs ~0.3% in CD; ~6% vs ~2% in UC) (Fig. 2).

Specifically, in our univariate analyses, the most significant association in CD is to HLA-DRB1*01:03 ($P<4\times10^{-62}$, OR= 2.53), with a p-value over 10 orders of magnitude more

significant than the next best associated variants in the region. Importantly, HLA-DRB1*01:03 has an effect in CD which is statistically independent from the other most associated variants in the MHC, as shown by reciprocal conditional logistic regression (**Supplementary** Fig. 1). In the UC univariate analysis the single most significant variant is a non-coding SNP (rs6927022, $P<5\times10^{-153}$, OR=1.49) near *HLA-DQA1*, previously identified in the recent meta-analysis of GWAS [4]; while multiple additional variants show highly significant association, most are correlated to this top signal (Fig. 1 and **Supplementary** Fig. 2). Strikingly the next strongest independent association is to HLA-DRB1*01:03, having a much greater OR ($P<1\times10^{-120}$, OR=3.63; $P_{cond}<2\times10^{-89}$, OR=3.06) (**Supplementary** Fig. 2). Reciprocal conditioning on HLA-DRB1*01:03 did not abolish the effect seen at rs6927022 ($P<9\times10^{-123}$, OR=1.43), indicating that these have mostly statistically independent effects in UC. Taken together, our analyses point to HLA-DRB1*01:03 as likely being causal in both diseases, with additional causal alleles in the class II and class I regions. Given this observation, it is probable that additional alleles within *HLA-DRB1* contribute to IBD risk.

We thus examined an *HLA-DRB1*-centric model and identified seven *HLA-DRB1* alleles with independent effects on CD risk (study-wide significance threshold of 5×10^{-6}) (**Supplementary Table 2**). Moreover, when controlling for these seven *HLA-DRB1* alleles, we identified only a single additional class II allele (HLA-DPA1*01:03) independently associated with CD. Using the same conditional logistic regression framework for the analysis of the class I locus, we identified seven class I HLA alleles that are significantly associated with CD, after conditioning on the eight class II alleles (Fig. 3 and **Supplementary Table 2**). This *HLA-DRB1*-centric model explains about 2% of disease variance (Fig. 2). In UC, we identified a total of 12 *HLA-DRB1* alleles, 1 *HLA-DPB1* allele and 3 class I alleles (**Supplementary Table 3**) that can explain the association to the MHC and which account for about 5% of disease variance (Fig. 2).

As can be seen in Figure 3, for many of the alleles identified in the *HLA-DRB1*-centric model, a few other candidate alleles in class I or class II can be considered. In particular, multiple HLA-DRB1 alleles have equivalent associations at *HLA-DQA1* and *HLA-DQB1* (e.g. HLA-DQA1*03:01 is equivalent to HLA-DRB1*04 and HLA-DRB1*09 alleles in UC) equally supporting a role for genetic variation within *HLA-DQA1* and/or *HLA-DQB1* in disease susceptibility, particularly for UC (Fig. 3). However, several of the alleles in these models, including HLA-DRB1*01:03, do not have any such proxies and thus are strong candidates for being causal (Fig. 3). Further dissection of these class II correlated signals for identifying potential causal alleles may only be feasible in admixed or ethnically diverse populations [9]. Further refinement may also be possible by examining the impact of clinical sub-phenotype and associated autoimmune co-morbidities on observed associations, although functional studies will be needed to infer causality. For the present analysis we were able to assess the impact of colonic vs. non-colonic inflammation, and found that HLA-DRB1*01:03 is associated with colonic CD and that HLA-DRB1*07:01 is associated with the absence of colon involvement (**Supplementary** Fig. 3), in line with previous suggestions[10]. This explains the shared associations for CD and UC at HLA-DRB1*01:03

and strongly suggests that this allele is critically involved in determining the colonic immune response to local flora.

Given that classical HLA alleles consist of combinations of specific amino acids at multiple positions, we tested whether the association to disease could be better explained by single amino acid positions. Indeed we observed very strong association signals at many single amino acid variants in CD (e.g. five amino acids of HLA-DRβ at positions 67, 70 and 71) and in UC (e.g. 4 amino acid variants of HLA-DQα at positions 50 and 53 and 215 and 4 amino acid variants in HLA-DRβ at positions 98 and 104) and also performed per position omnibus analyses that confirm the predominant association to HLA-DRβ position 11 in UC, as previously reported [5], and to HLA-DRβ position 70 in CD (**Supplementary Tables 4–5** and **Supplementary** Fig. 4). While the hypothesis of a positional effect is appealing, the interpretation of these position-based tests is not straightforward in the context of likely multiple causal alleles **(Supplementary Note on amino acids, Supplementary Table 6** and **Supplementary** Fig. 5). Furthermore in this study the amino acid-based models did not capture the association at *HLA-DRB*1 in a more parsimonious way than the HLA allele-based models (**Supplementary Note on amino acids**). To further explore the basis for the observed HLA associations, we performed three-dimensional protein structure modeling followed by analysis of the electrostatic properties of the binding groove of associated ($P<10^{-4}$) and common (frequency>1%) *HLA-DRB1* alleles. These analyses suggest that HLA-DR alleles associated with increased risk of UC and CD, share common structural and electrostatic properties within or near their peptide binding groove that are largely distinct from those of HLA-DR alleles associated with decreased risk of UC and CD (Fig. 4).

While we performed the primary analyses based on a dose effect model, our sample size allowed us to investigate the effects further, by testing for non-additive effects. In fact we found significant departure from additive effects in UC, but not for CD (Fig 5a–c). Specifically, we found evidence of decreased heterozygosity in UC patients for genotyped and imputed variants across the MHC and at HLA genes, mostly in class II (**Supplementary Tables 7–8**). This heterozygote advantage could be explained by an enrichment of dominant protective and recessive risk alleles[11], that is absent or much less important in CD (Fig. 5 and **Supplementary** Fig. 6). Notably, we also detected multiple overdominant effects in UC, the strongest of which is captured by HLA-DRB1*03:01 (Fig. 5, **Supplementary** Fig. 6–7, and **Supplementary Table 9)**. This allele is mostly found on the ancestral haplotype 8.1, a relatively common (~5–10%) haplotype that is conserved in European populations and that is implicated in other immune diseases [12–14]. The overdominance effect of this haplotype in UC is possibly due to the presence of both dominant protective and recessive risk alleles, which would be consistent with the reported recessive risk of this haplotype in the UC-related biliary disease primary sclerosing cholangitis (**Supplementary Fig. 8–9**) [15,16]. Analogous with an infectious paradigm [11], these data may suggest that decreased HLA class II heterozygosity may impair the ability to appropriately control colonic microbiota in UC.

Although there is a significant challenge in defining the causal alleles for CD and UC in the MHC given the LD structure in the region, a number of conclusions can be drawn regardless of the models tested. First, the high density mapping of this region in a large cohort revealed

the significant contribution of the MHC to disease risk, a contribution that is not apparent in the previous GWAS. Second, for both CD and UC it would appear that variation within HLA genes as opposed to variation in other genes within the MHC plays a predominant role in disease susceptibility. Third, while the contribution of class I and class II HLA variants to disease risk is relatively equivalent in CD, HLA class II variation plays a more important role in UC. Fourth, in contrast to the majority of non-MHC susceptibility loci being shared between CD and UC, most associated HLA alleles have a predominant role in either CD or UC, with very few having shared IBD risk Fig. 6). Finally, the decreased heterozygosity in UC suggests that the ability to recognize a broader set of antigens, potentially of colonic microbial origin, is important to mount protective immunity.

Online Methods

Genotype dataset

The cohorts used in the current study were collected from 15 countries across Europe, North America, and Australia, and have previously been described[4]. In total 19,802 CD cases, 14,864 UC cases and 34,872 controls of European ancestries, from the International IBD genetics consortium (IIBDGC; www.ibdgenetics.org), were included in the study.

Genotyping of the IIBDGC cohorts was performed in 34 different batches across 11 different genotyping centers, and additional genotyping data for 5,815 controls was obtained from the International MS Genetics Consortium (IMSGC)[17]. All participating centers received approval from their local and national institutional review boards, and informed consent was obtained from all participants in the study. All DNA samples included in the study were genotyped using the Immunochip custom genotyping array (Illumina, San Diego, California, USA)[4,18].

Genotype calling and quality control

After an initial genotype calling using Illumina's Bead Studio and a first stage of quality control, all data was centrally recalled using optiCall v.0.6.2[19]. OptiCall clustering was performed for each batch separately, with a Hardy-Weinberg Equilibrium (HWE)-threshold of 1×10^{-15}, HWE blanking disabled and a genotype call threshold of 0.7. HWE was calculated conditional on predicted ethnicity, and related individuals were removed from this calculation.

After recalling, a single unified QC procedure was performed across all genotyping batches, including the IMSGC controls. Variants that either failed the Hardy-Weinberg Equilibrium test in unaffected individuals, had different missing genotype rates in affected and unaffected individuals, or had significantly different allele frequencies across the batches based on false-discovery rate (FDR) threshold of 10^{-5} for each test were removed. We also removed variants that had missing genotype rate >2% across the entire collection, or > 10% in any single batch. Variants that only failed one QC criteria in a single batch were set to missing in the failed batch. Individuals were removed if they showed a missing genotype rate >2%, had a significantly higher or lower inbred coefficient (*F*) (plink)[20] at FDR <0.01, and showed a high level of relation (PI_HAT ≥ 0.4) calculated on IBS distance between all individuals. The coefficient of inbreeding and inter-sample relationships were calculated on

a LD-pruned dataset of independent variants. In order to control for population stratification while avoiding the possible bias introduced by the enrichment of associated alleles in the dataset, principal components were computed on control samples, based on a set of 18,123 independent (LD-pruned) SNPs across the Immunochip, and then applied to the affected samples. To generate the LD-pruned SNPs, we removed variants in long range LD and pruned the common variants (MAF>0.05) three times (plink[20]). Genomic inflation factor (λ) was estimated from a set of 3120 "null" SNPs (chosen based on GWAS of schizophrenia, psychosis and reading/mathematics ability), using different subset of principal components. Based on these and investigation of contribution of individual SNPs to the components (loadings), we selected to use the first five principal components to control for population stratification (**Supplementary Fig. 11**).

For the purpose of this study, the chr6:25Mb-34Mb region, encompassing MHC region, was extracted from the post-QC Immunochip dataset. In total, after QC, 18,405 CD cases, 14,308 UC and 34,241 healthy controls subjects were successfully genotyped for 8,001 SNPs within the MHC region.

Imputation of missing genotype data

Missing genotype data, from failed genotype calls or failed QC in single batches, were imputed using the Beagle SNP imputation package(v.3.0.4); imputation was performed using only the high density information contained within the dataset, no external reference dataset was used[21].

Imputation of HLA alleles

In order to avoid cohort specific asymmetry in the dataset, variants that failed QC in only one genotyping batch were removed before imputation of HLA alleles. Imputation of HLA alleles was done using two independent HLA imputation pipelines HLA*imp2[22] and SNP2HLA (v2)[23]; imputation of polymorphic amino acid positions and SNP variants was performed using SNP2HLA (v2). A set of additional SNP variants were also included using version 1 of SNP2HLA, which used a different reference dataset and imputed SNP variants not found in the SNP2HLA (v2) reference dataset.

Imputation accuracy

To benchmark HLA imputation results generated by HLA*IMP2 and SNP2HLA, we used two cohorts from the current study (Italian and Norwegian) for which classical HLA typing was available. The Italian dataset consisted of 450 ulcerative colitis cases and 280 controls for which 4-digit HLA types at *HLA-A, HLA-B, HLA-C, HLA-DQB1* and *HLA-DRB1* generated by Sequence-based typing (SBT)[24]. The Norwegian dataset contained 244 ulcerative colitis cases and 254 controls with 2-digit HLA types and a subset of 92 cases and 250 controls with 4-digit HLA types at HLA-*DRB1* generated at the Oslo University Hospital. We only considered individuals of whom both HLA alleles were successfully typed at 2- and 4-digit resolution at the locus under validation. For validation of HLA allele imputations, we compared imputation results of HLA*IMP2 and SNP2HLA to the lab-derived types in a locus-specific and allele-specific manner. We calculated per locus

concordance (**Supplementary Table 10**) and sensitivity, specificity, positive predictive value, negative predictive value and accuracy for each allele (**Supplementary Table 11**).

Let's denote true positives, true negatives, false positives and false negatives by *TP*, *TN*, *FP* and *FN*, respectively. Given our analyses were performed using expected allele doses from posterior probabilities, we used the expected doses for computation:

$$\begin{aligned} TP_i &= \min(x_i, y_i) \\ FP_i &= y_i - TP_i \\ TN_i &= \min(2-x_i, 2-y_i) \\ FN_i &= (2-y_i) - TN_i, \end{aligned}$$

where x_i and y_i are respectively the typed and imputed dose for individual, and the sum over all individuals gives the *TP*, *FP*, *TN* and *FN* (**Supplementary Table 11**). For a given HLA allele, we calculated sensitivity, specificity, positive predictive value, negative predictive value and accuracy using the usual definitions:

$$\begin{aligned} \text{sensitivity} &= TP/(TP+FN) \\ \text{specificity} &= TN/(TN+FP) \\ \text{positive predictive value } (PPV) &= TP/(TP+FP) \\ \text{negative predictive value } (NPV) &= TN/(TN+FN) \\ \text{accuracy} &= (TP+TN)/(TP+TN+FP+FN). \end{aligned}$$

For a given locus, we calculated the concordance as TP_{all}/(total number of chromosomes) (**Supplementary Table 10**).

Final dataset

As an additional QC step, we performed manual cluster inspection for any genotyped SNPs in the region tagging (r2>0.8) imputed SNPs, amino acid variants and HLA alleles reported in our proposed models for CD and UC.

In order to not duplicate HLA alleles in our dataset, and given the mostly equivalent imputation quality of the 2 pipelines, we opted to keep HLA imputations for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and *HLA-DPB1* derived from the SNP2HLA pipeline since amino acid and additional SNPs were also obtained from this pipeline, whereas the HLA allele information for the *HLA-DRB3*, *HLA-DRB4*, and *HLA-DRB5* genes were obtained from the HLA*imp2 pipeline.

Genotyped SNPs variants, as well as imputed SNPs, HLA alleles and amino acid variants at polymorphic amino acid positions were combined into a single dataset for analysis. Variants with a minor allele frequency of less than 0.05% in controls and variants showing imputation quality (INFO) score <0.5 were removed from the analyses. The final dataset contained 8,939 SNP variants, 138 4-digit resolution HLA alleles, 90 2-digit resolution HLA alleles and 741 single amino acid variants.

Heterogeneity

In order to evaluate heterogeneity of effects and allele frequencies between subgroups of different European ancestry, we clustered the individuals into relatively large population subgroups. These clusters were determined using k-means on the principal components. Based on the decline within cluster sum of squares, we determined that the optimal choice for the number of clusters (k) was in a range of 6–10. For each value of k in this range, we compared the clustering obtained to self-reported country of origin, when available. The objective was to be able to identify known population structures, while keeping homogeneous group in a single cluster. Based on these criteria, we stratified the dataset into 9 clusters (**Supplementary Fig. 12**). For every reported variant, we evaluated and illustrated heterogeneity of effect sizes between these 9 clusters using forest plots (Data on heterogeneity of effects is available on http://www.medgeni.org/goyette_nature_gen_2015).

Association testing and conditional analyses

Unless otherwise stated, all analyses were corrected for five principal components and performed in R (v 2.15.2) on expected allele counts (additive dose from posterior probability)(see **Supplementary Tables 4, 12 and 13** for primary univariate association results for single amino acid variants, SNP and HLA alleles respectively). Also, 4-digit HLA alleles were prioritized over 2-digit alleles in the final selection of association signals included in our models. We calculated the threshold for statistical significance in our study at $p<5\times10^{-6}$ for a study-wide type 1 error rate of 5% with 9,852 independent tests.

Given the imputed nature of the HLA allele data within our dataset, complexity of signals and burden of dimensionality, a standard forward conditional logistic regression approach was avoided. We opted instead to identify all HLA alleles showing study-wide significant association across the MHC region, in the primary univariate association dataset, and evaluate their independent effects through single pair-wise reciprocal conditional logistic regression (**Supplementary Tables 2–3**). This approach allowed the identification of independent association signals, composed either of single HLA alleles or groups of equivalent alleles (Fig. 3).

Gene based analyses (omnibus)

We tested the association to phenotypes at each HLA gene using logistic regression under an additive model of effect, including all 4-digit alleles with a frequency greater than 0.5%. Evidence of association at the gene is given as the p-value of the likelihood test for this regression model versus the null model (including only principal components). Evidence of association at a given gene, conditional on other HLA genes, is given as the p-value of the likelihood ratio test for the full model (including all HLA alleles) versus the partial model (without including the alleles at the given gene). In a similar fashion, we also tested the remaining association at each HLA gene, conditional on our final set of associated alleles (final model). To be noted, the different HLA genes have different number of distinct alleles, thus variable level of complexity (degrees of freedom) for the model and the interpretation of these gene-based tests is not straightforward in the context of likely multiple causal alleles (see the **Supplementary Note on amino acid**).

Variance explained (pseudo R^2)

In order to represent the importance of genetic variation in the MHC for CD and UC, we computed an estimate of variance explained for different models. Variance explained is not well defined for binary outcomes and many different metrics exist to represent it[25]. Given the correlation between the variants in the MHC, we computed McKelvey Zavoina's pseudo R^2 on the logit scale[26]. This metric can be computed using correlated variables and is independent of disease prevalence. Let β be the vector of fitted coefficients from logistic regression and S the estimated covariance matrix of the predictor, the McKelvey Zavoina's pseudo R^2 is then given by:

$$R^2_{MZ} = \frac{\beta' S \beta}{\beta' S \beta + \pi^{2/3}}.$$

To be noted, this estimation of variance explained shouldn't be directly compared to values given by other metrics or to heritability estimates based on Gaussian liability. We computed the variance explained by a regression model including all the HLA alleles with a frequency greater than 0.5%. We did the same separately for HLA alleles within class I and class II. Variance explained by class I, after inclusion of class II, was computed as the improvement in R^2 for the full model (all HLA alleles), compared to class II only. This difference in R^2 estimates the specific contribution of class I to the total variance explained that cannot be attributed to class II alleles. We did the same for class II. In order to be able to compare our results to the SNPs identified by the GWAS[4], we computed R^2 for the published GWAS index SNP in the MHC for CD and UC.

Subphenotype analyses

The IIBDGC has collected detailed subphenotype information for a subset of samples included in this study. These phenotypes include: demographics, disease location and behavior in CD, disease extent in UC, surgery and primary sclerosing cholangitis (PSC). Quality control of this subphenotype information and genotype-phenotype association testing has been performed in the context of another project from this IIBDGC (personal communication with Charlie Lee). In the context of this fine-mapping project, we considered disease location in CD and PSC in UC to evaluate the impact of disease heterogeneity. Association tests were performed within subset of samples with known subphenotype.

Non-additive effects, heterozygote advantage and overdominance

We tested evidence of non-additive effects at each variant using a logistic regression including terms for additive and non-additive effect, as described below. Rare variants with minor allele frequencies below 5% were excluded from this analysis, given that the very low number of homozygote precludes the evaluation of the model. For this particular analysis, we used best guess genotype data, unless otherwise stated. We computed evidence of non-additive effect as the p-value of the Wald statistic for the dominance term. We also computed the evidence of association for an allele under the general model using the likelihood ratio test.

Suppose a genetic variant with alleles *G* and *g*. The genotypes (GG, Gg, gg) can be coded as $u = (1, 0, -1)$ and $v = (0, 1, 0)$, respectively. In this context, u = dose - *1* and $v = 1 - |u|$. The effect of a specific genotype is then given as $au + dv$, where a and d are respectively the additive and dominance effect, as estimated by logistic regression. This parameterization can be generalized to expected allele counts (additive dose from posterior probability).

Under this parameterization, the effects of the genotypes *GG*, *Gg* and *gg* are given by a, d and $-a$, respectively. A strictly additive model would be $d = 0$ and $a \neq 0$, while a dominant or recessive model would be $d = a$ or $d = -a$. If the dominance term has a higher protective effect than the additive term, that is $d < -|a|$, the model is one of overdominance, where being a heterozygote provides protection, compared to both homozygotes.

For each HLA gene, we also tested for evidence of heterozygote advantage. We coded each individual as homozygote or heterozygote at the gene, as determined from the imputed 2-digit and 4-digit alleles. Association of phenotype and zygosity at each gene was tested using logistic regression. Evidence of association is given as the p-value of the likelihood test.

Pairwise comparisons of common 2-digit alleles at HLA-DRB1 were conducted to better understand the non-additive effects (**Supplementary Table 9**). For each pair or alleles, analysis was constrained to the subset of individual carrying only these two alleles as homozygotes or heterozygotes. Genotype effects were evaluated within each subset. Overdominance was tested as heterozygote vs the lowest risk homozygote. Heterozygote advantage was tested as heterozygote vs the pooled homozygote. Non-additive effect was tested based on the dominance term, as described previously.

Comparative structural modeling of HLA-DR alleles

Comparative structure models for all HLA-DR alleles associated with UC or CD, and of all common HLA-DR alleles (population frequency >1%), were generated using the program MODELLER[27]. Templates for comparative structure modeling were identified by querying the RCSB Protein Database[28] using the sequence of HLA-DRB1*01:01 and the DELTA-BLAST algorithm, for humans (Taxonomy ID: 9606, E-value threshold of 0.05). Of 36 templates identified, 10 crystallographically resolved HLA-DRB1 structures (PDB codes: 4MD5, 1FV1, 1KLU, 3PDO, 1D5M, 1JH8, 4MDI, 3L6F, 4I5B, 2IPK) were retained, which had high resolution (<2.5 Å) and favorable markers of structural quality (Ramachandran plot, DOPE, Verfiy3D and WHAT_CHECK scores)[29–32]. The sequences of the extracellular domain of target HLA molecules were retrieved from the EBI sequence database (ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/) and aligned using Clustal W2 using the *BLOSUM* matrix and the *Neighbor Joining* clustering algorithm and manually adjusted as indicated[33]. The modeled peptide was standardized to an alanine 12-mer, which was removed prior to calculating protein electrostatics to allow comparison of the electrostatic potential within the peptide-binding groove.

Electrostatic potential calculations

Atom charges and radii were assigned and side-chains protonated for pH 7.4 using the PARSE force field in PDB2PQR[34]. Protein electrostatic potential was calculated by solving the linearized Poisson-Boltzmann equation in APBS for a cubic grid of 353 points at a spacing of 0.33 Å[35]. Other parameters were set as follows: ionic solution of 0.15 M of univalent positive and negative ions; protein dielectric of 2; solvent dielectric of 78; temperature of 310 K; and a probe radius of 1.4 Å. HLA class II molecules typically make contact with nine amino acid residues of the presented peptide; of these, 7 peptide residues (at peptide positions 1, 2, 3, 4, 6, 7 and 9) make contact within the peptide binding groove[36]. Peptide residues at positions 5 and 8 are elevated away from the peptide-binding groove. For comparison of the electrostatic potential of the peptide-binding groove, radii were chosen so that the space around each coordinate encompassed all side chain atoms of the relevant peptide amino acid residue; the electrostatic potential within 1Å from the molecular surface of the HLA molecule was not examined.

The peptides of the template HLA structures were used to define the geometric average coordinate of the positions of the side chain atoms of peptide amino acid residues 1, 2, 3, 4, 6, 7 and 9. The electrostatic potential within a 3.5–5 Å radius from each coordinate was considered for comparison of the electrostatic properties of the peptide-binding groove among HLA-DR alleles (**Supplementary Fig. 10**). For comparison of the electrostatic potential around amino acid positions 67, 70 and 71, (associated to both UC and CD in different analyses; **Supplementary Tables 4–6**), the geometric average coordinate of the positions of these amino acids' side chain atoms was calculated and the electrostatic potential within a 5 Å radius from this coordinate was considered for comparisons among different alleles. Electrostatic potential comparisons were performed using the Hodgkin's index as described previously[27,28] in a pairwise, all-versus-all, fashion to produce a distance matrix[37,38]. Distance matrices were displayed as a symmetrical heatmap with re-ordering such that electrostatically similar alleles are clustered together, according to the dendrogram (**Supplementary Fig. 13**). A pooled heatmap was created using the Euclidian distance between the individual distance matrices from the 7 peptide-binding groove regions (Fig. 4a). Heatmaps were created for electrostatic potential comparisons at individual regions of the HLA-DR molecule (7 regions in the peptide-binding groove and one region defined by amino acid residues 67, 70 and 71, as detailed above) (Fig. 4b).

Acknowledgments

J.D.R. holds a Canada Research Chair and his current work is supported by grants from the U.S. National Institute of Diabetes and Digestive and Kidney Diseases (DK064869; DK062432). A.F.'s lab is supported by the German Ministry of Education and Research (BMBF) grant program e:Med (sysINFLAME). A.F. receives infrastructure support from the Deutsche Forschungsgemeinschaft (DFG) Cluster of Excellence "Inflammation at Interfaces" and holds an endowment professorship (Peter Hans Hofschneider Professorship) of the Foundation for Experimental Biomedicine (Zurich, Switzerland). Grant support for T.H.K. and A.F. was received from EU 7th Framework Programme (FP7/2007-2013, grant number 262055, ESGI). This project has been funded in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research (M.N.C). J.B. was supported by a Wellcome Trust grant (#WT098051). D.H.M. and V.K. are supported by the NIHR Cambridge Biomedical Research Centre. P.S. is supported by the NIDDK IBD Genetics Consortium Data Coordinating Center grant (7

U01 DK062429-14). J.T. is supported by the Medical Research Council. D.P.B.M. is supported by The Leona M. and Harry B. Helmsley Charitable Trust, The European Union (305479), and grants DK062413, DK046763-19, AI067068, HS021747 and U54DE023789-01. R.H.D holds the Inflammatory Bowel Disease Genetic Research endowed chair at the University of Pittsburgh and was supported by an NIDDK IBD Genetics Consortium Genetic Research Center grant (DK062420) and a U.S. National Cancer Institute grant (CA141743). S.L.H. and J.R.O. would like to also acknowledge the support of National Institutes of Health (R01 NS049477; 1U19A1067152), and the National Multiple Sclerosis Society (RG 2899-D11). S.L. wishes to acknowledge the support from the Australian National Health and Medical Research Council (RD Wright Career Development Fellowship, APP1053756). We would like to thank the International PSC study group (www.ipscsg.org) for sharing data. We are grateful to Benedicte A. Lie and Kristian Holm for helpful discussions.

References

1. Horton R, et al. Gene map of the extended human MHC. Nat Rev Genet. 2004; 5:889–99. [PubMed: 15573121]
2. Rioux JD, et al. Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. Proc Natl Acad Sci U S A. 2009; 106:18680–5. [PubMed: 19846760]
3. Stokkers PC, Reitsma PH, Tytgat GN, van Deventer SJ. HLA-DR and -DQ phenotypes in inflammatory bowel disease: a meta-analysis. Gut. 1999; 45:395–401. [PubMed: 10446108]
4. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–24. [PubMed: 23128233]
5. Achkar JP, et al. Amino acid position 11 of HLA-DRbeta1 is a major determinant of chromosome 6p association with ulcerative colitis. Genes Immun. 2012; 13:245–52. [PubMed: 22170232]
6. Jones DC, et al. Killer Ig-like receptor (KIR) genotype and HLA ligand combinations in ulcerative colitis susceptibility. Genes Immun. 2006; 7:576–82. [PubMed: 16929347]
7. Kulkarni S, et al. Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease. Proc Natl Acad Sci U S A. 2013; 110:20705–10. [PubMed: 24248364]
8. Satsangi J, et al. Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease. Lancet. 1996; 347:1212–7. [PubMed: 8622450]
9. Oksenberg JR, et al. Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. Am J Hum Genet. 2004; 74:160–7. [PubMed: 14669136]
10. Newman B, et al. CARD15 and HLA DRB1 alleles influence susceptibility and disease localization in Crohn's disease. Am J Gastroenterol. 2004; 99:306–15. [PubMed: 15046222]
11. Lipsitch M, Bergstrom CT, Antia R. Effect of human leukocyte antigen heterozygosity on infectious disease outcome: the need for allele-specific measures. BMC Med Genet. 2003; 4:2. [PubMed: 12542841]
12. Alper CA, Fleischnick E, Awdeh Z, Katz AJ, Yunis EJ. Extended major histocompatibility complex haplotypes in patients with gluten-sensitive enteropathy. J Clin Invest. 1987; 79:251–6. [PubMed: 3793924]
13. Aly TA, et al. Multi-SNP analysis of MHC region: remarkable conservation of HLA-A1-B8-DR3 haplotype. Diabetes. 2006; 55:1265–9. [PubMed: 16644681]
14. Wiencke K, Spurkland A, Schrumpf E, Boberg KM. Primary sclerosing cholangitis is associated to an extended B8-DR3 haplotype including particular MICA and MICB alleles. Hepatology. 2001; 34:625–30. [PubMed: 11584356]
15. Donaldson PT, Norris S. Evaluation of the role of MHC class II alleles, haplotypes and selected amino acid sequences in primary sclerosing cholangitis. Autoimmunity. 2002; 35:555–64. [PubMed: 12765483]
16. Liu JZ, et al. Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. Nat Genet. 2013; 45:670–5. [PubMed: 23603763]
17. International Multiple Sclerosis Genetics C et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet. 2013; 45:1353–60. [PubMed: 24076602]
18. Trynka G, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet. 2011; 43:1193–201. [PubMed: 22057235]
19. Shah TS, et al. optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. Bioinformatics. 2012; 28:1598–603. [PubMed: 22500001]

20. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–75. [PubMed: 17701901]
21. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009; 84:210–23. [PubMed: 19200528]
22. Dilthey A, et al. Multi-population classical HLA type imputation. PLoS Comput Biol. 2013; 9:e1002877. [PubMed: 23459081]
23. Jia X, et al. Imputing amino acid polymorphisms in human leukocyte antigens. PLoS One. 2013; 8:e64683. [PubMed: 23762245]
24. Gourraud PAKP, Cereb N, Yang SY, Feolo M, Maiers M, Rioux JD, Hauser S, Oksenberg J. HLA diversity in the 1000 Genomes dataset. PloS One. 2014
25. Veall MRZKF. Pseudo-R2 measures for some common limited dependent variable models. J econometric surveys. 10:241–259.
26. McKelvey RDZW. A statistical model for the analysis of ordinal level dependent variables. J Math Sociology. 1975; 4:103–120.
27. Eswar N, et al. Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci. 2007; Chapter 2(Unit 2):9. [PubMed: 18429317]
28. Berman HM, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–42. [PubMed: 10592235]
29. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature. 1996; 381:272. [PubMed: 8692262]
30. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993; 26:283–291.
31. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature. 1992; 356:83–5. [PubMed: 1538787]
32. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006; 15:2507–24. [PubMed: 17075131]
33. Larkin MA, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23:2947–8. [PubMed: 17846036]
34. Dolinsky TJ, et al. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res. 2007; 35:W522–5. [PubMed: 17488841]
35. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A. 2001; 98:10037–41. [PubMed: 11517324]
36. Jones EY, Fugger L, Strominger JL, Siebold C. MHC class II proteins and disease: a structural perspective. Nat Rev Immunol. 2006; 6:271–82. [PubMed: 16557259]
37. Richter S, Wenzel A, Stein M, Gabdoulline RR, Wade RC. webPIPSA: a web server for the comparison of protein interaction properties. Nucleic Acids Res. 2008; 36:W276–80. [PubMed: 18420653]
38. Wade RC, Gabdoulline RR, De Rienzo F. Protein interaction property similarity analysis. Int J Quantum Chem. 2001; 83:122–127.

Members of the International Inflammatory Bowel Disease Genetics Consortium

Clara Abraham[30], Jean-Paul Achkar[31,32], Tariq Ahmad[33], Leila Amininejad[34,35], Ashwin N Ananthakrishnan[36,37], Vibeke Andersen[38,39], Carl A Anderson[19], Jane M Andrews[40], Vito Annese[11,12], Guy Aumais29,[41], Leonard Baidoo[17], Robert N Baldassano[42], Tobias Balschun[4], Peter A Bampton[43], Murray Barclay[44], Jeffrey C Barrett[19], Theodore M Bayless[45], Johannes Bethge[46], Joshua C Bis[47], Alain Bitton[48], Gabrielle Boucher[1], Stephan Brand[49], Steven R Brant[45], Carsten Büning[50], Angela Chew[51,52], Judy H Cho[53], Isabelle Cleynen[54], Ariella Cohain[55], Anthony Croft[56], Mark J Daly[7,8], Mauro D'Amato[57], Silvio Danese[58], Dirk De Jong[59], Martine De Vos[60], Goda Denapiene[61], Lee A Denson[62], Kathy L Devaney[36], Olivier Dewit[63], Renata D'Inca[64], Marla Dubinsky[65], Richard H Duerr[17,18], Cathryn Edwards[66], David Ellinghaus[4], Jonah Essers[67,68], Lynnette R Ferguson[69], Eleonora A Festen[70], Philip Fleshner[20], Tim Florin[71], Denis Franchimont[34,35], Andre

[30]Section of Digestive Diseases, Department of Internal Medicine, Yale School of Medicine, NewHaven, Connecticut, USA.
[31]Department of Gastroenterology and Hepatology, Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio, USA.
[32]Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA.
[33]Peninsula College of Medicine and Dentistry, Exeter, UK.
[34]Department of Gastroenterology, Erasmus Hospital, Brussels, Belgium.
[35]Department of Gastroenterology, Free University of Brussels, Brussels, Belgium.
[36]Gastroenterology Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[37]Division of Medical Sciences, Harvard Medical School, Boston, Massachusetts, USA.
[38]Medical Department, Viborg Regional Hospital, Viborg, Denmark.
[39]Organ Center, Hospital of Southern Jutland Aabenraa, Aabenraa, Denmark.
[40]Inflammatory Bowel Disease Service, Department of Gastroenterology and Hepatology, Royal Adelaide Hospital, Adelaide, Australia.
[41]Department of Gastroenterology, Hôpital Maisonneuve-Rosemont, Montréal, Québec, Canada.
[42]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.
[43]Department of Gastroenterology and Hepatology, Flinders Medical Centre and School of Medicine, Flinders University, Adelaide, Australia.
[44]Department of Medicine, University of Otago, Christchurch, New Zealand.
[45]Meyerhoff Inflammatory Bowel Disease Center, Department of medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
[46]Department for General Internal Medicine, Christian-Albrechts-University, Kiel, Germany.
[47]Cardiovascular Health Research Unit, University of Washington, Seattle, Washington, USA.
[48]Division of Gastroenterology, Royal Victoria Hospital, Montréal, Québec, Canada.
[49]Department of Medicine II, Ludwig-Maximilians-University Hospital Munich-Grosshadern, Munich, Germany.
[50]Department of Gastroenterology, Campus Charité Mitte, Universitatsmedizin Berlin, Berlin, Germany.
[51]IBD unit, Fremantle Hospital, Fremantle, Australia.
[52]School of Medicine and Pharmacology, University of Western Australia, Fremantle, Australia.
[53]Department of Genetics, Yale School of Medicine, New Haven, Connecticut, USA.
[54]Department of Clinical and experimental medicine, Translational Research in GastroIntestinal Disorders (TARGID), Katholieke Universiteit (KU) Leuven, Leuven, Belgium.
[55]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA.
[56]Inflammatory Bowel Diseases, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia.
[57]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden.
[58]IBD Center, Department of Gastroenterology, Istituto Clinico Humanitas, Milan, Italy.
[59]Department of Gastroenterology and Hepatology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands.
[60]Department of Hepatology and Gastroenterology, Ghent University Hospital, Ghent, Belgium.
[61]Center of hepatology, Gastroenterology and Dietetics, Vilnius University, Vilnius, Lithuania.
[62]Pediatric Gastroenterology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA.
[63]Department of Gastroenterology, Université Catholique de Louvain (UCL) Cliniques Universitaires Saint-Luc, Brussels, Belgium.
[64]Division of Gastroenterology, University Hospital Padua, Padua, Italy.
[65]Department of Pediatrics, Cedars Sinai Medical Center, Los Angeles, California, USA.
[66]Department of Gastroenterology, Torbay Hospital, Torbay, Devon, UK.
[67]Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[68]Pediatrics, Harvard Medical School, Boston, Massachusetts, USA.
[69]Faculty of Medical & Health Sciences, School of Medical Sciences, The University of Auckland, Auckland, New Zealand.
[70]Department of Gastroenterology and Hepatology, University Medical Center Groningen, Groningen, The Netherlands.

Franke[4], Karin Fransen[72], Richard Gearry[44,73], Michel Georges[74,75], Christian Gieger[76], Jürgen Glas[48], Philippe Goyette[1], Todd Green8,[67], Anne M Griffiths[77], Stephen L Guthery[78], Hakon Hakonarson[42], Jonas Halfvarson[79,80], Katherine Hanigan[56], Talin Haritunians[20], Ailsa Hart[81], Chris Hawkey[82], Nicholas K Hayward[83], Matija Hedl[30], Paul Henderson[84,85], Xinli Hu[86], Hailiang Huang[7,8], Ken Y Hui[53], Marcin Imielinski[42], Andrew Ippoliti[20], Laimas Jonaitis[87], Luke Jostins[5,6], Tom H Karlsen[26,27,28], Nicholas A Kennedy[88], Mohammed Azam Khan[89,90], Gediminas Kiudelis[87], Subra Kugathasan[91], Limas Kupcinskas[92], Anna Latiano[11], Debby Laukens[60], Ian C Lawrance[52], James C Lee[93], Charlie W Lees[88], Marcis Leja[94], Johan Van Limbergen[95], Paolo Lionetti[96], Jimmy Z Liu[19], Edouard Louis[97], Gillian Mahy[98], John Mansfield[99], Dunecan Massey[93], Christopher G Mathew[100,101], Dermot PB McGovern[20], Raquel Milgrom[102], Mitja Mitrovic[72,103], Grant W Montgomery[83], Craig Mowat[104], William Newman[89,90], Aylwin Ng[36,105], Siew C Ng[106], Sok Meng Evelyn Ng[30], Susanna Nikolaus[46], Kaida Ning[30], Markus Nöthen[107], Ioannis Oikonomou[30], Orazio Palmieri[11], Miles Parkes[93], Anne Phillips[104], Cyriel Y Ponsioen[108], Urõs Potocnik[103,109], Natalie J Prescott[100,101], Deborah D Proctor[110], Graham Radford-Smith[111,112], Jean-Francois Rahier[113], Soumya

[71]Department of Gastroenterology, Mater Health Services, Brisbane, Australia.
[72]Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands.
[73]Department of Gastroenterology, Christchurch Hospital, Christchurch, New Zealand.
[74]Unit of Animal Genomics, Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA-R) Research Center, University of Liege, Liege, Belgium.
[75]Faculty of Veterinary Medicine, University of Liege, Liege, Belgium.
[76]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany.
[77]Gastroenterology, Hepatology and Nutrition, The Hospital for Sick Children, Toronto, Ontario, Canada.
[78]Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, Utah, USA.
[79]Department of Medicine, Örebro University Hospital, Örebro, Sweden.
[80]School of Health and Medical Sciences, Örebro University, Örebro, Sweden.
[81]Department of Medicine, St Mark's Hospital, Harrow, Middlesex, UK.
[82]Nottingham Digestive Diseases Centre, Queens Medical Centre, Nottingham, UK.
[83]Genetic Epidemiology, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia.
[84]Paediatric Gastroenterology and Nutrition, Royal Hospital for Sick Children, Edinburgh, UK.
[85]Child Life and Health, University of Edinburgh, Edinburgh, Scotland, UK.
[86]Division of Rheumatology Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA.
[87]Academy of Medicine, Lithuanian University of Health Sciences, Kaunas, Lithuania.
[88]Gastrointestinal Unit, Wester General Hospital University of Edinburgh, Edinburgh, UK.
[89]Genetic Medicine, Manchester Academic Health Science Centre, Manchester, UK.
[90]The Manchester Centre for Genomic Medicine, University of Manchester, Manchester, UK.
[91]Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA.
[92]Department of Gastroenterology, Kaunas University of Medicine, Kaunas, Lithuania.
[93]Inflammatory Bowel Disease Research Group, Addenbrooke's Hospital, Cambridge, UK.
[94]Faculty of medicine, University of Latvia, Riga, Latvia.
[95]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Hospital for Sick Children, Toronto, Ontario, Canada.
[96]Dipartimento di Neuroscienze, Psicologia, Area del Farmaco e Salute del Bambino (NEUROFARBA), Università di Firenze SOD Gastroenterologia e Nutrizione Ospedale pediatrico Meyer, Firenze, Italy.
[97]Division of Gastroenterology, University Hospital CHU of Liege, Liege, Belgium.
[98]Department of Gastroenterology, The Townsville Hospital, Townsville, Australia.
[99]Institute of Human Genetics, Newcastle University, Newcastle upon Tyne, UK.
[100]Department of Medical and Molecular Genetics, Guy's Hospital, London, UK.
[101]Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London, UK.
[102]Inflammatory Bowel Disease Centre, Mount Sinai Hospital, Toronto, Ontario, Canada.
[103]Center for Human Molecular Genetics and Pharmacogenomics, Faculty of Medicine, University of Maribor, Maribor, Slovenia.
[104]Department of Medicine, Ninewells Hospital and Medical School, Dundee, UK.
[105]Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[106]Department of Medicine and Therapeutics, Institute of Digestive Disease, Chinese University of Hong Kong, Hong Kong.
[107]Department of Genomics Life & Brain Center, University Hospital Bonn, Bonn, Germany.
[108]Department of Gastroenterology, Academic Medical Center, Amsterdam, The Netherlands.
[109]Faculty for Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia.

Raychaudhuri[86], Miguel Regueiro[17], Florian Rieder[31], John D Rioux[1,29], Stephan Ripke[7,8], Rebecca Roberts[44], Richard K Russell[84], Jeremy D Sanderson[114], Miquel Sans[115], Jack Satsangi[88], Eric E Schadt[55], Stefan Schreiber4,[46], L Philip Schumm[21], Regan Scott[17], Mark Seielstad[116,117], Yashoda Sharma[30], Mark S Silverberg[118], Lisa A Simms[111], Jurgita Skieceviciene[87], Sarah L Spain[101], A. Hillary Steinhart[102], Joanne M Stempak[102], Laura Stronati[119], Jurgita Sventoraityte[92], Stephan R Targan[20], Kirstin M Taylor[114], Anje ter Velde[108], Emilie Theatre[74,75], Leif Torkvist[120], Mark Tremelling[121], Andrea van der Meulen[122], Suzanne van Sommeren[70], Eric Vasiliauskas[20], Severine Vermeire[54,123], Hein W Verspaget[122], Thomas Walters[77,124], Kai Wang[42], Ming-Hsi Wang[31,45], Rinse K Weersma[70], Zhi Wei[125], David Whiteman[126], Cisca Wijmenga[72], David C Wilson[84,85], Juliane Winkelmann[127,128], Ramnik J Xavier8,[36], Sebastian Zeissig[46], Bin Zhang[55], Clarence K Zhang[129], Hu Zhang[130,131], Wei Zhang[30], Hongyu Zhao[129], Zhen Z Zhao[83], Australia and New Zealand IBDGC, Belgium IBD Genetics Consortium, Italian Group for IBD Genetic Consortium, NIDDK Inflammatory Bowel Disease Genetics Consortium, Quebec IBD Genetics Consortium, United Kingdom IBDGC, Wellcome Trust Case Control Consortium

[110]Section of Digestive Diseases, Department of Medicine, Yale University, New Haven, Connecticut, USA.
[111]Inflammatory Bowel Diseases, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia.
[112]Department of Gastroenterology, Royal Brisbane and Womens Hospital, Brisbane, Australia.
[113]Department of Gastroenterology, Université Catholique de Louvain (UCL) Centre hospitalier (CHU) Mont-Godinne, Mont-Godinne, Belgium.
[114]Department of Gastroenterology, Guy's & St Thomas' NHS Foundation Trust, St-Thomas Hospital, London, UK.
[115]Department of Digestive Diseases, Hospital Quiron Teknon, Barcelona, Spain.
[116]Human Genetics, Genome Institute of Singapore, Singapore.
[117]Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA.
[118]Inflammatory Bowel Disease Centre, Mount Sinai Hospital, Toronto, Ontario, Canada.
[119]Department of Biology of Radiations and Human Health, Agenzia nazionale per le nuove tecnologie l'energia e lo sviluppo economico sostenibile (ENEA), Rome, Italy.
[120]Department of Clinical Science Intervention and Technology, Karolinska Institutet, Stockholm, Sweden.
[121]Gastroenterology & General Medicine, Norfolk and Norwich University Hospital, Norwich, UK.
[122]Department of Gastroenterology, Leiden University Medical Center, Leiden, The Netherlands.
[123]Division of Gastroenterology, University Hospital Gasthuisberg, Leuven, Belgium.
[124]Faculty of medicine, University of Toronto, Toronto, Ontario, Canada.
[125]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA.
[126]Molecular Epidemiology, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia.
[127]Institute of Human Genetics, Technische Universität München, Munich, Germany.
[128]Department of Neurology, Technische Universität München, Munich, Germany.
[129]Department of Biostatistics, School of Public Health, Yale University, NewHaven, Connecticut, USA.
[130]Department of Gastroenterology, West China Hospital, Chengdu, Sichuan, China.
[131]State Key Laboratory of Biotherapy, Sichuan University West China University of Medical Sciences (WCUMS), Chengdu, Sichuan, China

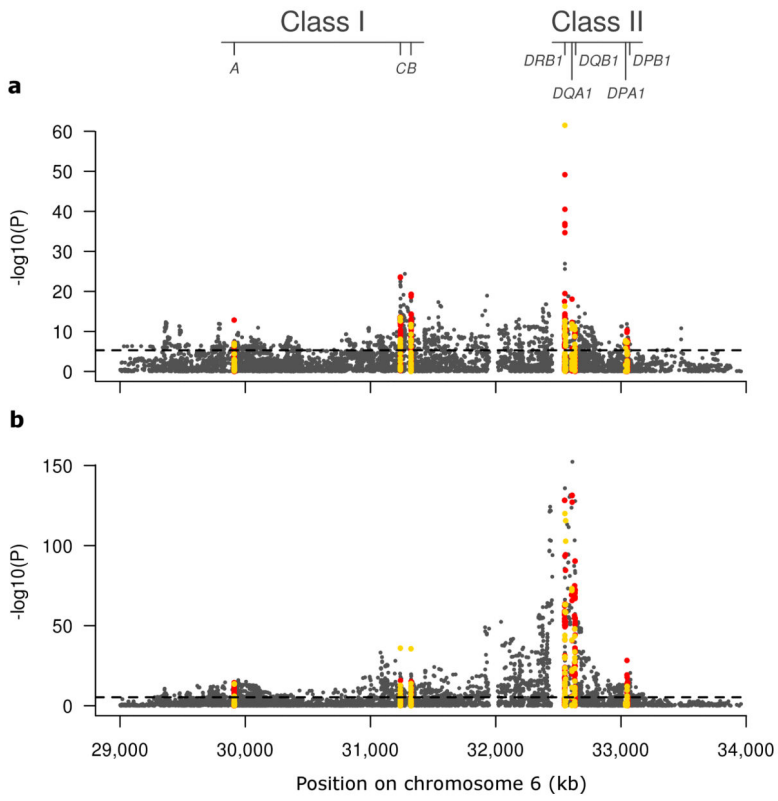

Figure 1. Primary univariate association analyses of CD and UC

Univariate association analysis results for 8,939 SNPs (dark grey) **(Supplementary Table 12)**, 90 2-digit and 138 4-digit resolution HLA alleles (yellow) (**Supplementary Table 13**), as well as 741 single amino acid variants (red) (**Supplementary Table 4**) in the MHC region are shown for 18,405 CD cases and 14,308 UC cases (with 34,241 common control subjects). Given that previous genetic analyses have identified distinct effects in the MHC for CD and UC, with different non-correlated alleles identified in each disease, we opted to perform the finemapping analyses for CD and UC separately. (**a**) The primary univariate association analysis in CD reveals over 1,789 markers showing study-wide significant association ($P<5\times10^{-6}$) across the MHC, including 32 4-digit resolution classical HLA alleles (Fig. 3 and **Supplementary Table 2**). The single most significant variant for CD is HLA-DRB1*01:03 ($P=3\times10^{-62}$, OR= 2.51). (**b**) The primary univariate association analysis in UC reveals over 2,762 markers showing study-wide significant association across the MHC, including 50 4-digit resolution classical HLA alleles (Fig. 3 and **Supplementary Table 3**). The single most significant variant for UC is rs6927022 ($P=8\times10^{-154}$, OR= 1.49) while the best HLA allele is HLA-DRB1*01:03 ($P=3\times10^{-119}$, OR=3.59); each acting independently. Twenty-nine SNPs and 9 amino acid variants surpass HLA-DRB1*01:03 as the next most significant variants in the primary analysis however all of these are correlated to rs6927022 and their significance is dramatically reduced by conditional logistic regression.

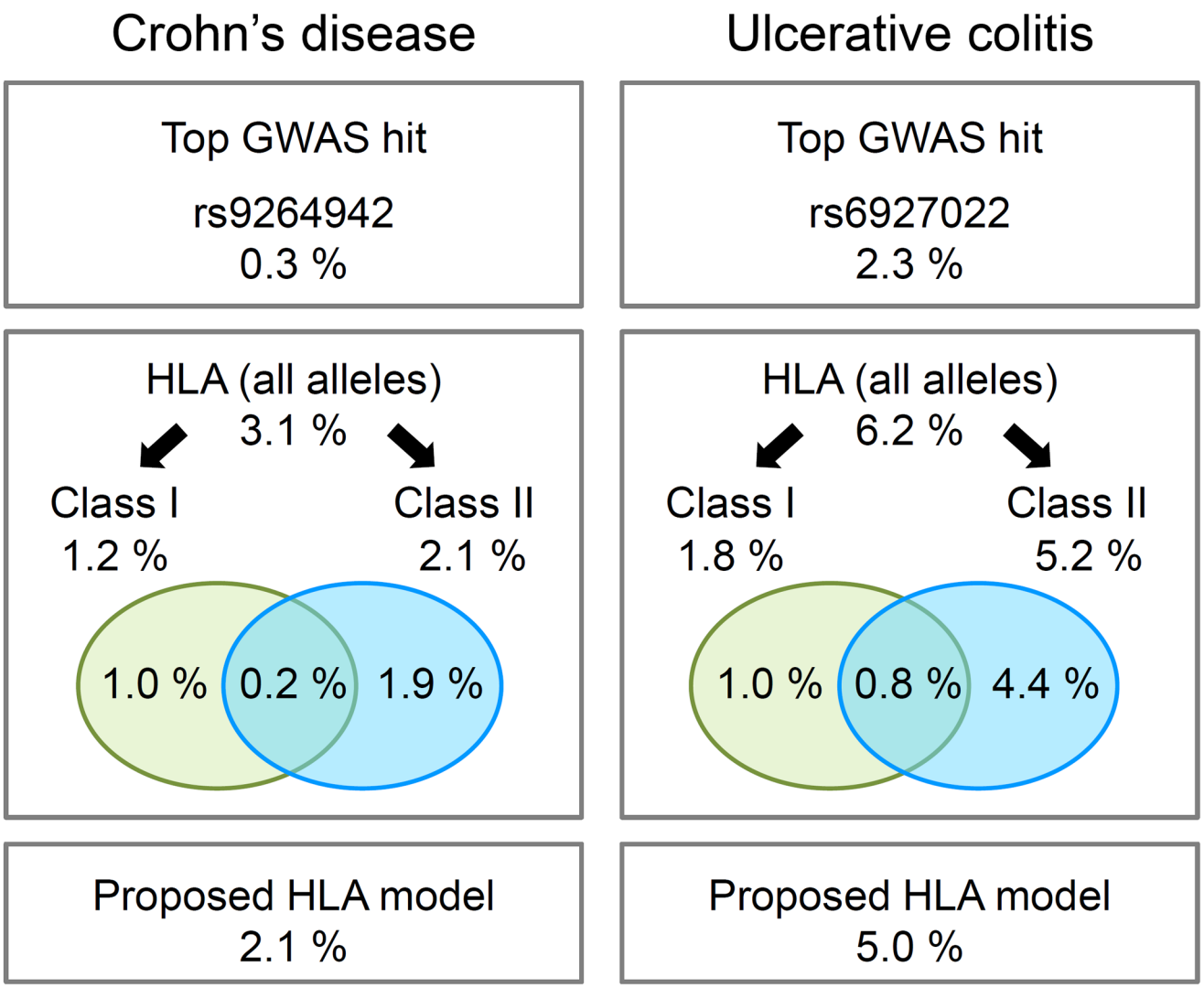

Figure 2. Variance explained by 4-digit HLA alleles in CD and UC

Proportion of variance explained on a logit scale (McKelvey and Zavoina's Pseudo R^2, see **Online Methods**) for different models in CD (left) and UC (right). The top boxes show the variance explained by previously identified GWAS index SNPs within the MHC[4]. The middle boxes illustrate the variance explained by HLA models including all 4-digit alleles of frequency > 0.5% (126 alleles in CD and UC) and models restricted to 4-digit alleles within either class I (63 alleles) or class II regions (63 alleles), respectively. The Venn diagram illustrates the proportion of variance explained that is unique to class I, class II or shared. The bottom boxes indicate the variance explained by the proposed HLA models (15 and 16 alleles in CD and UC, respectively). To be noted, these estimations of variance explained were performed on the logit scale for practical reasons, and should not be directly compared to heritability estimates computed on the (Gaussian) liability scale.

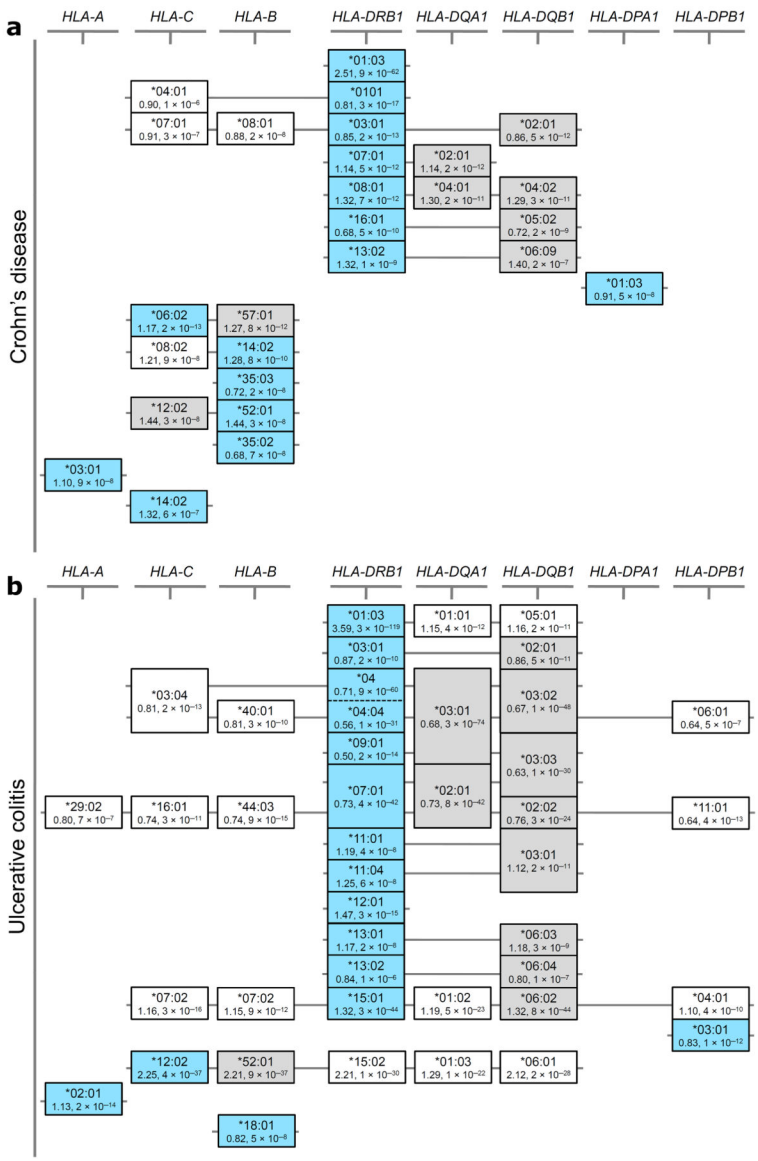

Figure 3. Correlated association signals at HLA alleles support potential alternate association models for both CD and UC

Equivalence of effect at the different study-wide significant associated 4-digit HLA alleles is shown for **(a)** CD and **(b)** UC. The structures illustrated in the figure are not classically defined haplotype structures, but were identified entirely based on the correlation of signal defined through pairwise reciprocal conditional logistic regression analyses (see **Supplementary tables 2 and 3**); although such correlations are clearly dependent on the underlying haplotypic structure of the region. Alleles identified as primary tags for independent association signals in our *HLA-DRB1* focused models are shown in light blue boxes, while alternate alleles with equivalent effects are shown in grey boxes. Alleles in white boxes show study-wide significant secondary effects that can be explained entirely by the selected HLA alleles. Alleles at the *HLA-DRB3, -DRB4 and -DRB5* genes were omitted in order to simplify the display; many of the alleles at these genes show high frequency and

as such are correlated to many different alleles (both risk and protective) at the other class II genes. Of note, the HLA-DRB4*null allele is the second strongest associated allele in UC **(see Supplementary table 3).**

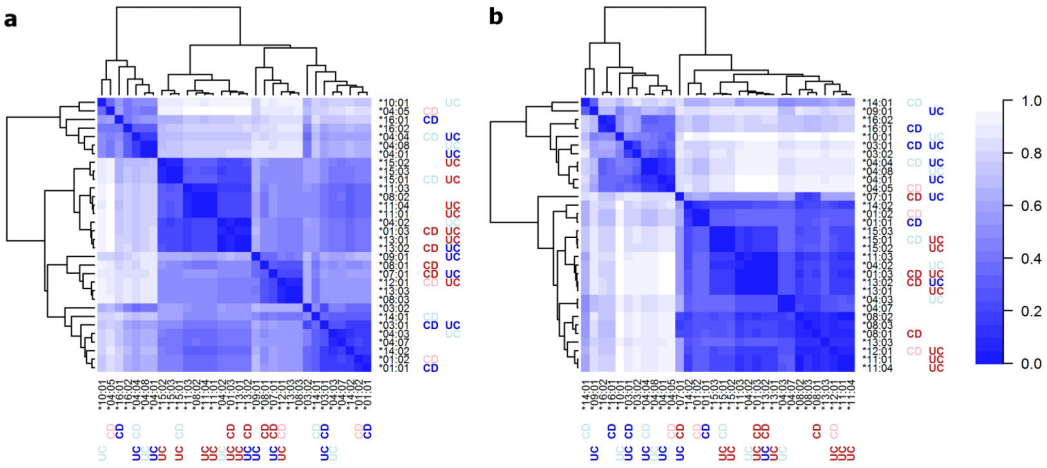

Figure 4. HLA-DR peptide binding groove electrostatic properties and risk of IBD

The electrostatic potential of all HLA-DR alleles associated with UC or CD, and of all common HLA-DR alleles (frequency >1%), was calculated. HLA-DR alleles associated with increased or decreased risk of IBD at study wide-significance level ($P< 5\times10^{-6}$) are shown in dark red or dark blue, respectively. Respective risk associations at suggestive level ($1\times10^{-4}<P<5\times10^{-6}$) are shown in pale red and pale blue. Electrostatic potential comparisons among HLA-DR molecules were performed in a pairwise, all-versus-all, fashion (see **Online Methods**) to produce distance matrices that are displayed as symmetrical heatmaps (scale ranges from 0 [identical] to 1 [maximum difference]). (**a**) The electrostatic potential in seven regions within the peptide binding groove (see **Online Methods** and **Supplementary Fig. 10**), which interact with the presented peptide, were compared among the HLA-DR alleles and pooled onto a single Euclidian distance matrix. The distance-based clustering identifies four clusters, with an enrichment of risk alleles in two of these. Comparison of the electrostatic potential at individual peptide binding groove regions is shown in **Supplementary Fig. 13**. (**b**) Heatmap representing electrostatic potential differences among the HLA-DR alleles at a spherical region that encompasses amino acid residues 67, 70 and 71 of the HLA-DRβ chain (associated with risk for UC and CD; **Supplementary Table 13**). The distance-based clustering identifies two clusters that correlate with directionality of effect in IBD.

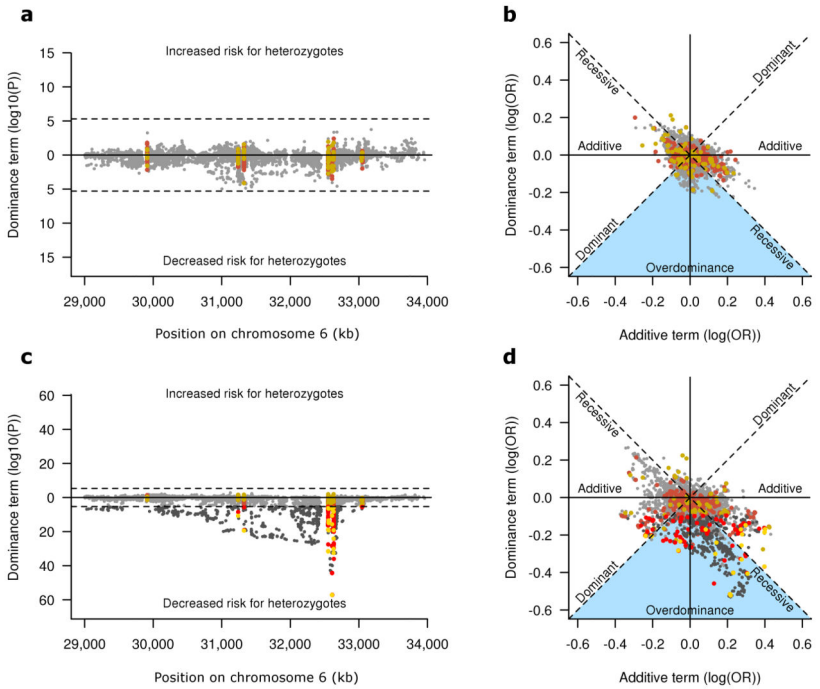

Figure 5. Non-additive effect models in CD and UC

Evidence for non-additive effect of common variants (frequency >5%) across the MHC tested under a general model of additive and dominance effects (**Online Methods**) in CD (**a,b**) and UC (**c,d**). The p-values and directionality for departure from additive effect (dominance term) are represented on the y-axis (**a,c**). HLA alleles and amino acids variants are in yellow and red respectively, while SNPs are represented in dark grey. Variants with non-significant ($P>5\times10^{-6}$) dominance term are plotted in less pronounced colors. A clear enrichment for lower risk in heterozygotes is observed in UC (**c**) as suggested by the large number of significant negative dominance term (lower part of the plot). This effect is absent in CD (**a**), or much less important. The dominance term OR is illustrated (y-axis) versus the additive term (x-axis) (**b,d**). Protective and risk minor alleles are shown on the left and right sides of the plot respectively. Strictly recessive or dominant variants are expected to fall on the diagonals, while strictly additive variants lay on or close to the x-axis. The y-axis is the expected position for pure over/under dominance. In UC (**d**), many alleles fall into the region of the plot for protective dominant, risk recessive or overdominance (blue triangle) (see **Supplementary Table 9** for pairwise comparison of HLA-DRB1 alleles). These non-additive effects are observed for many variants in UC (**c,d**) (e.g HLA-DRB1*03:01 and HLA-DQB1*02:01) but are mostly absent in CD (**a,b**); notable exception being the HLA-B*08 allele (**Supplementary** Fig. 6).

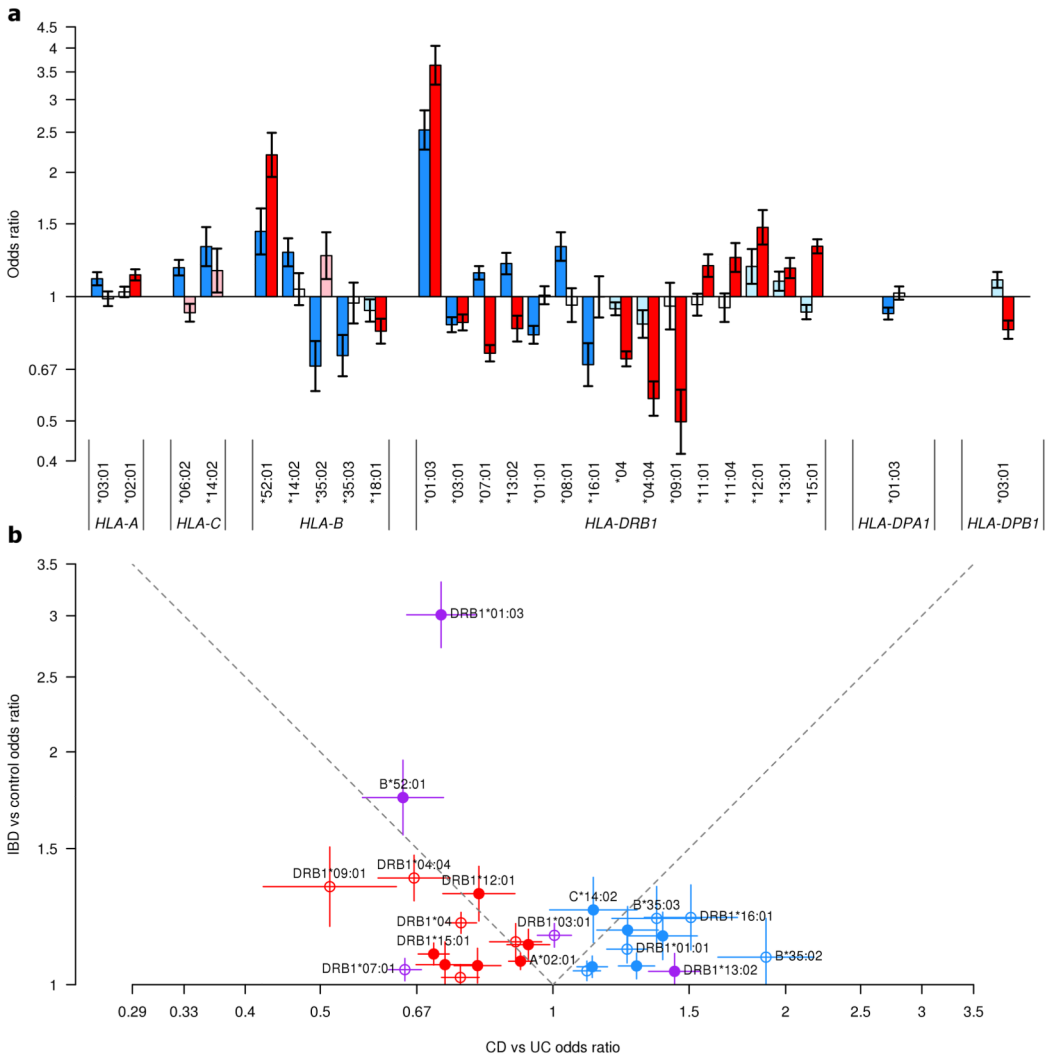

Figure 6. Comparison of odds ratio in CD and UC for HLA alleles identified from HLA-focused models

Odds ratio (OR) from the primary univariate association analyses in CD and UC for all alleles identified in the HLA-focused models of CD and/or UC are presented with 95% confidence intervals (**a**). Odds ratio for CD and UC are in blue and red respectively; darker colors indicate study-wide significant effect ($P<5\times10^{-6}$), lighter colors indicate nominal significance level ($0.05>P\ 5\times10^{-6}$) and white indicates non-significance (P 0.05) (for specific effect and significance values refer to Fig. 3 and **Supplementary Tables 2 and 3**). Allele HLA-B*52:01 is indicated for UC in place of the equivalent HLA-C*12:02 to simplify the display of this shared signal. For the same HLA alleles, odds ratio (with 95% confidence intervals) for an IBD analysis are plotted against the odds ratio for the CD versus UC analysis with the IBD risk allele as the reference (**b**). Empty circles represent variants where the absence of the allele is the reference. Alleles identified as significant in CD or UC only are plotted in blue and red, respectively. Variants identified as significant in both are shown in purple. To be noted, HLA-DRB1*07:01 and HLA-DRB1*13:02 have opposite direction of effect between CD and UC. Shared association signals are expected to fall in the

upper triangle of the plot. Most variants fall outside of this region, highlighting the difference between CD and UC in the MHC.